논문 2010-47SP-1-21

음성 부재 확률을 이용한 음성 강화 이득 수정 기법

(Robust Speech Reinforcement Based on Gain-Modification incorporating Speech Absence Probability)

최 재 훈*, 장 준 혁**

(Jae-Hun Choi and Joon-Hyuk Chang)

요 약

본 논문에서는 배경 잡음 환경에서 배경 잡음에 의해 저하된 음성 신호의 명료도를 soft decision 기반의 음성 부재 확률을 이용하여 음성 강화 이득을 수정함으로써 음성의 명료도를 보다 향상시키는 기법을 제안한다. 배경 잡음 환경에서 저하된 음성의 명료도를 향상시키기 위한 기존의 음성 강화 기법으로써 soft decision을 이용하여 오염된 음성 신호로부터 깨끗한 음성 신호만 증폭시키는 알고리즘이 제안되었다. 기존의 음성 강화 기법 보다 음성 구간과 비음성 구간 및 전이 구간에서 강인한 음성 강화 이득을 추정하기 위하여 soft decision 기반의 음성 부재 확률 (Speech Absence Probability)을 음성 강화 이득에 통합한 음성 강화 이득 수정 알고리즘을 제안한다. 제안된 음성 강화 기법의 성능은 다양한 배경 잡음 환경에서 ITU-T P.800 의 주관적인 음질 측정 방법인 (Comparison Category Rating) 테스트에 의해서 평가되었으며, 기존의 음성 강화 기법과 비교하여 향상된 성능을 보여주었다.

Abstract

In this paper, we propose a robust speech reinforcement technique to enhance the intelligibility of the degraded speech signal under the ambient noise environments based on soft decision scheme incorporating a speech absence probability (SAP) with speech reinforcement gains. Since the ambient noise significantly decreases the intelligibility of the speech signal, the speech reinforcement approach to amplify the estimated clean speech signal from the background noise environments for improving the intelligibility and clarity of the corrupted speech signal was proposed. In order to estimate the robust reinforcement gain rather than the conventional speech reinforcement method between speech active periods and nonspeech periods or transient intervals, we propose the speech reinforcement algorithm based on soft decision applying the SAP to the estimation of speech reinforcement gains. The performances of the proposed algorithm are evaluated by the Comparison Category Rating (CCR) of the measurement for subjective determination of transmission quality in ITU-T P.800 under various ambient noise environments and show better performances compared with the conventional method.

Keywords: Soft Decision, 음성강화, SNR 복구 기법, 음성 부재 확률

I. 서 론

다양한 배경 잡음 환경에서 상대방과 휴대폰을 통하여 통화하는 경우, 주변의 에워싼 배경 잡음으로 인해서 음성의 명료도 및 음질은 매우 감소된다. 이러한 배경 잡음으로 인해 음성의 명료도가 저하되는 문제를 해결하기 위한 방법으로써, 음성 향상 기법이 제안되었다[1~5].

음성 향상 기법은 근단으로 음성 신호를 전송하기 전에 원단의 잡음 환경에서 잡음에 의해 오염된 음성 신호의 잡음 부분을 줄이는 방법으로 잡음 제거에 초점을 맞춘다. 그러나 오염된 음성 신호로부터 잡음이 제거된 음성 신호가 근단에 전송되어 근단 화자가 듣게 되는

* 학생회원, ** 정회원, 인하대학교 전자공학부
(Department of Electronics Engineering, Inha University)

※ 본 논문은 2009년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2009-0085162). 그리고, 본 연구는 지식경제부와 한국산업기술재단의 전략기술인력양성사업으로 수행된 연구결과임.
접수일자: 2009년5월18일, 수정완료일: 2009년12월28일

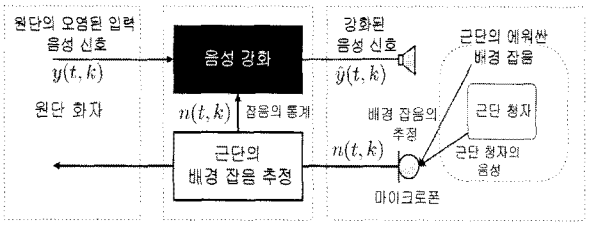

그림 1. 추정된 근단 화자의 배경 잡음 전력에 기반한 제안된 음성 강화 시스템 개략도

Fig. 1. Block diagram of the proposed speech reinforcement system based on the estimated near-end listener's background noise spectrum.

경우, 근단 화자의 환경이 잡음 환경이라면, 근단의 배경 잡음은 근단 화자의 귀에 직접적으로 영향을 미치게 되므로 잡음 제거된 음성 신호가 전송될지라도 배경 잡음에 의해 근단 화자가 듣게 되는 명료도 역시 저하되게 된다. 또한 이러한 배경 잡음은 직접적으로 제어되기 어렵기 때문에, 근단 화자를 에워싼 주변의 잡음 환경에서 근단 화자를 위한 원단 음성 신호의 강화 기법이 필요하다.

음성 강화 기법이란 주변의 잡음에 따라 상대 쪽으로부터 전송되어 온 음성을 증폭시킴으로써 청취 음성의 명료성을 개선하는 방법이다[6]. 이러한 배경 잡음에 대한 연구는 활발히 진행되고 있으며, 특히 근단의 배경 잡음 환경에서 근단 화자가 듣게 되는 음성의 명료도를 향상시키기 위한 최신의 연구로써 인간 청각의 loudness (인간이 느끼는 소리의 크기) 지각 모델을 이용한 음성 강화 방법, Sauert와 Vary의 주파수별 SNR 복구 기법, 그리고 soft decision에 기반한 음성 강화 기법을 들 수 있다[6~8].

인간 청각의 라우드니스 지각 모델을 이용한 음성 강화 방법은 주파수 별로 잡음이 없는 상태에서 전송되어 온 음성 신호가 들렸을 크기와 동일한 크기로 느껴지도록 주변 잡음에 따라 전송된 음성 신호를 증폭시켜주는 방법이다. 또한 SNR 복구 기법은 주파수 밴드별로 신호 대 잡음 비 (Signal-to-Noise Ratio, SNR)가 일정하게 유지되도록 전송되어 온 원단 음성 신호를 증폭시키는 기법이다[6]. 즉, 근단의 추정된 배경 잡음 전력에 기반하여 원단으로부터 전송되어 온 음성 신호의 모든 주파수 성분에 동일한 이득으로 증폭시킴으로써 음질 및 음성의 명료도를 향상시키는 방법이다. SNR 복구 기법의 경우 원단 화자의 환경이 잡음이 없는 환경이라는 점과 원단 음성 신호가 잡음이 섞이지 않은 깨끗한 음성 신호라 가정하기 때문에, 일반적인 잡음이 존재하는 통신 환경을 고려한다면, 실제 상황에 적용하기에는 상당한 문제점을 가지게 된다. 즉, 오염된 원단 음성 신호가 근단의 입력으로 들어오게 되는 경우, 음성 신호뿐 아니라 잡음 신호도 증폭시키기 때문에, 근단 화자가 듣게 되는 음성의 명료도는 더욱 저하되게 된다[7].

음성 신호뿐만 아니라 잡음 신호도 증폭시키는 문제점을 해결하기 위한 방법으로써, soft decision에 기반하여 오염된 음성 신호로부터 깨끗한 음성 신호만을 추정하여 증폭시키는 음성 강화 기법이 제안되었다[8]. 그러나 soft decision에 기반한 음성 강화 기법은 주파수 성분별 음성 부재 확률 (Speech Absence Probability: SAP)을 이용하여 오염된 음성 신호로부터 깨끗한 음성 신호를 추정하여 증폭시키게 되는데, 이 기법은 음성 구간과 비음성 구간 및 음성 구간에서 비 음성 구간으로의 전이 구간에서 음성 강화 이득의 추정에 있어서 강인하지 못한 문제점을 가지게 된다.

본 논문에서는, 추정된 배경 잡음 전력에 기반하여, 원단의 오염된 음성 신호 중에서 잡음이 섞이지 않은 깨끗한 음성 신호 추정 및 음성 강화 이득의 보다 강인한 추정을 위하여 soft decision 기반의 음성 부재 확률을 이용한 원단 음성 강화 이득 수정 알고리즘을 제안한다. 먼저, 주파수별 음성 부재 확률을 적용하여, 오염된 원단 음성 신호로부터 잡음이 섞이지 않은 깨끗한 음성만을 추정하고, 추정된 깨끗한 음성 신호와 근단으로부터 추정된 배경 잡음 전력과의 연산을 통하여 강화 이득을 추정하였다. 추정된 강화 이득에 음성 부재 확률 (SAP)을 결합하여, 음성 구간 및 비 음성 구간, 그리고 음성 구간에서 비음성 구간으로의 전이 구간에서 음성의 존재 유무의 확률 값에 따라 잡음이 섞이지 않은 깨끗한 음성 신호만을 증폭시킴으로써, 근단의 배경 잡음으로 인해 저하된 오염된 원단 음성 신호의 명료도를 향상시키는 기법을 제안한다. 제안된 알고리즘의 성능은 다양한 근단 배경 잡음 환경을 고려하여, ITU-T P.800의 주관적 음질 측정 방법인 CCR (Comparison Category Rating) 테스트로 실험을 진행하였으며, 기존의 SNR 복구 기법 및 주파수별 주파수 성분에 soft decision을 적용한 방법과 비교하여 보다 향상된 결과를 나타내었다[9].

본 논문의 구성은 다음과 같다. 먼저 Ⅱ장에서는 기존의 음성 강화 기법에 이해를 위하여 주파수별 SNR 복구 기법과 Soft decision 기반 음성 강화 기법에 대하여

서술하고, Ⅲ장에서는 제안된 음성 부재 확률을 이용한 음성 강화 이득 수정 기법을 기술하였다. Ⅳ장에서는 실험 결과 비교 및 분석을 기술하였으며, 마지막 Ⅴ장에서는 결론을 맺는다.

Ⅱ. 기존의 음성 강화 기법의 이해

1. 주파수별 SNR 복구 기법의 개요

배경 잡음이 음성 신호에 미치는 영향에 대한 연구로써, Sauert와 Vary에 의해 근단 화자를 위한 원단 음성 신호의 명료도를 향상시키는 알고리즘이 제안되었다[7]. 주파수별 SNR 복구 기법은 모든 주파수 밴드에 대해 일정한 SNR을 유지하도록 각각의 주파수 성분을 증폭시키는 방법이다. 먼저, 시간축 상에서 오염된 원단 음성 신호와 잡음이 섞이지 않은 깨끗한 음성 신호, 근단의 배경 잡음 신호를 각각 $y(t)$와 $s(t)$, $n(t)$로 각각 나타낸다면, 이산 푸리에 변환 (Discrete Fourier Transform: DFT)을 통한 주파수 축 변환에 의해 각각 $Y(t,k)$와 $S(t,k)$, $N(t,k)$로 나타낼 수 있다. 여기서 $Y(t,k)$와 $S(t,k)$ 그리고 $N(t,k)$는 t번째 프레임에서의 k번째 주파수 성분을 나타낸다. 음성 강화 이득 $G(t,k)$로 원단 음성 신호의 DFT 계수를 증폭한다면, 증폭된 원단의 음성 신호는 다음과 같이 표현된다.

$$\hat{Y}(t,k) = G(t,k)\,Y(t,k) \tag{1}$$

기존의 SNR 복구 기법에 따르면, $\Phi_{\hat{Y}\hat{Y}}(t,k)$와 $\Phi_{NN}(t,k)$의 신호 대 잡음 비 (SNR)는 특정한 SNR보다 다 크거나 커야 하며, 다음과 같이 표현된다[7].

$$\frac{\Phi_{\hat{Y}\hat{Y}}(t,k)}{\Phi_{NN}(t,k)} \geq \xi \tag{2}$$

식 (2)에서 $\Phi_{YY}(t,k)$와 $\Phi_{\hat{Y}\hat{Y}}(t,k)$ 그리고 $\Phi_{NN}(t,k)$은 각각 원단 음성 신호와 증폭된 음성 신호 및 근단 배경 잡음 신호의 short-term power density를 나타낸다. 식 (2)는 다음과 같은 식으로 나타낼 수 있다.

$$\frac{\Phi_{\hat{Y}\hat{Y}}(t,k)}{\Phi_{NN}(t,k)} = \frac{G^2(t,k)\Phi_{YY}(t,k)}{\Phi_{NN}(t,k)} \geq \xi \tag{3}$$

또한, 식 (3)을 증폭 이득 값 $G(t,k)$에 관한 식으로 나타내면 다음과 같다.

$$G(t,k) \geq \sqrt{\xi \frac{\Phi_{NN}(t,k)}{\Phi_{YY}(t,k)}} \tag{4}$$

원단 음성 신호는 근단의 배경 잡음 환경에서 감쇄되지 않아야 하기 때문에 다음과 같은 최소 증폭 이득 값은 다음과 같이 도입할 수 있다.

$$G(t,k) \geq 1 \tag{5}$$

따라서, 식 (4)와 식 (5)의 최소 증폭 이득 값을 결합하면 다음의 식과 같이 나타낼 수 있다.

$$G(t,k) = \max\left\{\sqrt{\xi \frac{\Phi_{NN}(t,k)}{\Phi_{YY}(t,k)}}, 1\right\} \tag{6}$$

다양한 배경 잡음 환경에서 근단 화자가 듣게 되는 원단 음성 신호가 과도하게 증폭되지 않도록 $G(t,k)$의 최대 증폭 이득 조건을 $G_{\max} \cong 30dB$로 제한한다. 따라서 식 (6)과 최대 증폭 이득 조건을 결합하면, 최종적으로 다음과 같은 강화 이득을 구할 수 있다.

$$G(t,k) = \min\left\{\max\left\{\sqrt{\xi \frac{\Phi_{NN}(t,k)}{\Phi_{YY}(t,k)}}, 1\right\}, G_{\max}\right\} \tag{7}$$

식 (6)에서 나타난 원단 입력 음성 신호의 short-term PSD $\Phi_{YY}(t,k)$와 근단의 배경 잡음 신호의 short-term PSD $\Phi_{NN}(t,k)$은 다음과 같이 추정되어진다.

$$\begin{aligned}\Phi_{YY}(t,k) &= \alpha_Y \Phi_{YY}(t-1,k) + (1-\alpha_Y)|Y(t,k)|^2 \\ \Phi_{NN}(t,k) &= \alpha_N \Phi_{NN}(t-1,k) + (1-\alpha_N)|N(t,k)|^2\end{aligned} \tag{8}$$

원단 입력 음성 신호의 short-term PSD와 근단의 배경 잡음 신호의 short-term PSD를 추정함에 있어, 실험적으로 최적화된 고정 시간 상수 $\alpha_Y = 0.996$와 $\alpha_N = 0.96$을 사용하였다[7].

2. Soft Decision에 기반 음성 강화 기법의 개요

기존의 주파수별 SNR 복구 기법의 경우, 각각의 주파수 성분에 적용되는 강화 이득 값은 현재 프레임에 대한 원단 음성 신호 전력 $|Y(t,k)|$에 따라 결정된다. 더욱이, 식 (7)에처럼, 강화 이득 $G(t,k)$은 $\Phi_{YY}(t,k)$와 $\Phi_{NN}(t,k)$의 함수로써 표현될 수 있기 때문에, 원단의 음성 신호가 원단의 배경 잡음에 의해 오염된 음성

신호라면, 음성 신호 뿐만 아니라 부가된 잡음 신호도 증폭되어 질 수 있다. 음성 신호와 잡음 신호가 모두 증폭된다면, 음성의 명료도가 오히려 더 저하되기 때문에, 오염된 원단 음성 신호로부터 잡음이 섞이지 않은 깨끗한 음성 신호만을 추정하여 , 추정된 깨끗한 음성 신호만을 증폭시키는 것이 음성의 명료도 향상에 무엇보다 중요하다. 따라서 원단의 오염된 음성 신호로부터 잡음이 섞이지 않은 깨끗한 음성 신호만을 추정하는 것이 필요하다. 오염된 원단의 음성 신호로부터 잡음이 섞이지 않은 깨끗한 음성 신호만을 추정하기 위하여, Soft Decision에 기반한 음성 강화 기법이 제안되었[8]. 구체적으로, 원단 음성 신호에 대한 각각의 주파수 밴드에서의 음성 존재 및 음성 부재를 분류하기 위하여 다음 식과 같이 가정할 수 있다.

$$\begin{aligned} &speech\ absence\ H_0 : Y(t,k) = D(t,k) \\ &speech\ presence\ H_1 : Y(t,k) = S(t,k) + D(t,k) \end{aligned} \tag{9}$$

식 (9)에서 $S(t,k)$와 $D(t,k)$은 원단 음성 신호에 대해서 깨끗한 음성과 부가된 잡음에 대한 각각의 DFT 계수를 나타낸다.

음성 신호와 잡음 신호의 스펙트럼이 zero-mean 복소 가우시안 분포를 갖는다고 가정하면, 확률 밀도 함수는 음성 부재 및 음성 존재에 대한 가설 H_1, H_0에 따라 표현 될 수 있다[1].

$$\begin{aligned} p(Y(t,k)|H_0) &= \frac{1}{\pi\lambda_d(t,k)} exp\left\{ - \frac{\Phi_{YY}(t,k)}{\lambda_d(t,k)} \right\} \\ p(Y(t,k)|H_1) &= \frac{1}{\pi\left[\lambda_d(t,k)+\lambda_s(t,k)\right]} \\ &\quad \cdot \exp\left\{ - \frac{\Phi_{YY}(t,k)}{\lambda_d(t,k)+\lambda_s(t,k)} \right\} \end{aligned} \tag{10}$$

식 (10)에서 $\lambda_s(t,k)$와 $\lambda_d(t,k)$은 각각 t번째 프레임에서의 k번째 주파수 성분에 대한 음성과 잡음의 분산을 의미한다. 따라서 음성 존재 및 부재에 관한 가설로부터 주파수 채널별 음성 부재 확률 (Speech Absence probability: SAP) $p(H_0|Y(t,k))$은 다음과 같이 주어진다[1].

$$\begin{aligned} p(H_0|Y(t,k)) &= \frac{p(Y(t,k)|H_0)p(H_0)}{p(Y(t,k))} \\ &= \frac{p(Y(t,k)|H_0)p(H_0)}{p(Y(t,k)|H_0)p(H_0)+p(Y(t,k)|H_1)p(H_1)} \\ &= \frac{1}{1+\frac{p(H_1)}{p(H_0)}\Lambda(Y(t,k))} \end{aligned} \tag{11}$$

위의 식에서, $p(H_0)=(1-p(H_1))$은 음성 부재에 대한 사전 확률 (a priori probability)이고, $\Lambda(Y(t,k))$는 k번째 주파수 대역에서의 우도비 (likelihood ratio)로써 다음 식에 의해 표현된다.

$$\begin{aligned} \Lambda(Y(t,k)) &= \frac{p(Y(t,k)|H_1)}{p(Y(t,k)|H_0)} \\ &= \frac{1}{1+\xi(t,k)} exp\left[\frac{\gamma(t,k)\xi(t,k)}{1+\xi(t,k)} \right] \end{aligned} \tag{12}$$

여기서, $\gamma(t,k)$와 $\xi(t,k)$은 각각 predicted SNR과 a posteriori SNR를 의미하며, 다음과 같이 나타내어진다.

$$\begin{aligned} \gamma(t,k) &\equiv \frac{\Phi_{YY}(t,k)}{\Phi_{DD}(t,k)} \\ \xi(t,k) &\equiv \frac{\Phi_{SS}(t,k)}{\Phi_{DD}(t,k)} \end{aligned} \tag{13}$$

제안된 음성 강화 기법에 있어서, 오염된 원단 음성 신호로부터 잡음이 섞이지 않은 깨끗한 음성 신호만을 추정하여 증폭하는 것이기 때문에, 잡음 전력 $\Phi_{DD}(t,k)$과 잡음이 섞이지 않은 깨끗한 음성 전력 $\Phi_{SS}(t,k)$의 추정이 무엇보다 중요하다. 일반적으로, 음성 검출기 (Voice activity detector:VAD)를 사용하여 음성 부재 구간에서 잡음 전력을 갱신한다[1]. 그러나 잡음 환경이 비정상 (non-stationary)이라면, 잡음 전력은 음성의 존재 구간에서도 갱신되어야 한다. 따라서 음성 전력 $\Phi_{SS}(t,k)$와 잡음 전력 $\Phi_{DD}(t,k)$을 각각 스무딩 (smoothing)을 사용하여 다음과 같이 나타낸다.

$$\begin{aligned} \hat{\Phi}_{SS}(t+1,k) &= \zeta_s \hat{\Phi}_{SS}(t,k) + (1-\zeta_s)\hat{\Phi}_{SS}(t,k) \\ \hat{\Phi}_{DD}(t+1,k) &= \zeta_d \hat{\Phi}_{DD}(t,k) + (1-\zeta_d)\hat{\Phi}_{DD}(t,k) \end{aligned} \tag{14}$$

식 (14)에서, $\hat{\Phi}_{DD}(t,k)$와 $\hat{\Phi}_{SS}(t,k)$는 잡음 전력 $\Phi_{DD}(t,k)$와 음성 전력 $\Phi_{SS}(t,k)$의 추정 값이고, $\zeta_s(=0.98)$와 $\zeta_d(=0.99)$는 정상 (stationary) 상태를 가정한 스무딩 파라미터로써 $0<\zeta_s,\zeta_d<1$의 값을 가

진다. 주파수 채널별 음성 부재 확률 $p(H_0|Y(t,k))$와 음성 신호 $S(t)$와 잡음 신호 $D(t)$의 통계적 가정을 이용해서 현재 프레임에 대한 음성 신호 및 잡음 신호의 추정 값을 계산하면 다음과 같다.

$$\begin{aligned}\hat{\Phi}_{SS}(t,k) = & E\left[|S(t,k)|^2|Y(t,k), H_0\right]p(H_0|Y(t,k)) \\ & + E\left[|S(t,k)|^2|Y(t,k), H_1\right]p(H_1|Y(t,k))\end{aligned} \tag{15}$$

$$\begin{aligned}\hat{\Phi}_{DD}(t,k) = & E\left[|D(t,k)|^2|Y(t,k), H_0\right]p(H_0|Y(t,k)) \\ & + E\left[|D(t,k)|^2|Y(t,k), H_1\right]p(H_1|Y(t,k))\end{aligned} \tag{16}$$

따라서 식 (7)과 식 (8) 그리고 현재 프레임에서 추정된 잡음이 섞이지 않은 깨끗한 음성 신호 전력 $\hat{\Phi}_{SS}(t,k)$을 SNR 복구 기법에 통합하면 새로운 음성 강화 이득 $G(t,k)$를 구할 수 있다.

$$G(t,k) = \min\left\{\sqrt{\xi\frac{\hat{\Phi}_{NN}(t,k)}{\hat{\Phi}_{SS}(t,k)}}, G_{\max}\right\} \tag{17}$$

여기서, 식 (7)에서의 $\Phi_{YY}(t,k)$은 제안된 음성 강화 알고리즘에 의해 잡음이 섞이지 않은 깨끗한 음성 신호만을 추정한 $\hat{\Phi}_{SS}(t,k)$으로 바뀌게 된다.

III. Soft Decision을 이용한 음성 강화 이득 수정 기법

오염된 원단 음성 신호로부터 깨끗한 음성 신호를 추정한 후, 추정된 잡음이 섞이지 않은 깨끗한 음성 신호만을 증폭시키기 위해서는 강화 이득 값 $G(t,k)$의 추정이 무엇보다 중요하다. 강화 이득 값 $G(t,k)$에 soft decision을 통합함으로써, 음성의 존재 확률에 따라 음성 강화가 이루어지도록 하였다. 구체적으로, 강화 이득 값 $G(t,k)$에 음성 부재 확률 $p(H_0|Y(t,k))$을 통합하여 다음과 같은 새로운 강화 이득 값 $\tilde{G}(t,k)$을 정의하였다.

$$\begin{aligned}\tilde{G}(t,k) = & G(t,k)(1-p(H_0|Y(t,k)) \\ & + G_{\min}p(H_0|Y(t,k))\end{aligned} \tag{18}$$

식 (18)에서 $G_{\min}(=1)$은 강화 이득 값은 최소 값을 나타내게 된다. 음성이 존재하지 않는 구간에서는 강화 이득 값 $G(t,k)$를 1로 함으로써 잡음이 증폭되어 음성의 명료도가 저하되는 것을 방지한다. 또한 음성 구간에서 비음성 구간으로의 전이 구간 $(0 < p(H_0|Y(t,k)) < 1)$에서는 $G_{\min}$과 $G(t,k)$를 결합함으로써 보다 강인한 강화 이득 값을 추정이 가능하도록 하였다. 따라서 음성 부재 확률에 따라 강화 이득 값은 다음과 같이 정리된다.

$$\tilde{G}_{total}(t,k) = \begin{cases} G(t,k), & p(H_0|Y(t,k)) \approx 0 \\ \tilde{G}(t,k), & 0 < p(H_0|Y(t,k)) < 1 \\ G_{\min}(t,k), & p(H_0|Y(t,k)) \approx 1 \end{cases} \tag{19}$$

식 (19)에서 볼 수 있는 것처럼, 수정된 강화 이득은 기존의 오염된 음성 신호로부터 깨끗한 음성 신호를 추정하기 위하여 soft decision을 적용한 음성 강화 기법과 비교하여 음성과 비음성의 전이 구간에서 보다 강인한 깨끗한 음성 신호 증폭이 이루어지도록 한다는 점에서 큰 차이를 보임을 알 수 있다[8]. 그림 2에서는 제안된 soft decision 기반의 음성 강화 이득 수정 기법의 전체 블록도를 나타내었다.

제안된 soft decision 통합 음성 강화 이득 수정 알고리즘이 기존의 오염된 음성 신호로부터 깨끗한 음성 신호만을 추정하여 증폭시키는 음성 강화 기법과 비교하여 더 강인한 성능을 보인다는 것은 그림 3을 통해서 확인할 수 있다. 그림 3을 보면 (a)은 오염된 원단 음성 신호에 대한 음성 부재 확률 (SAP)값을 나타내고, (b)에는 기존의 음성 강화 기법으로 구한 강화 이득 값을 나타낸다. 또한 (c)은 제안된 soft decision 통합 음성 강화 이득 수정 알고리즘에 의해서 구해진 음성 강화 이득 값을 보여준다. 그림에서 보는 바와 같이, 음서 부재 확률이 1에 가까운 구간, 즉 음성이 거의 없다고 판

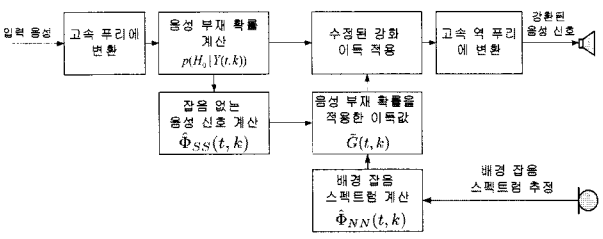

그림 2. 제안된 음성 강화 이득 수정 기법의 전체 블록도

Fig. 2. Overall block diagram of the proposed speech reinforcement approach.

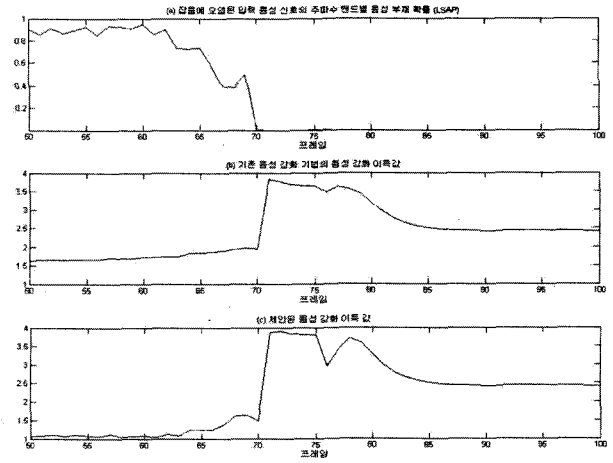

그림 3. 음성 강화 이득 값의 비교. (a) 오염된 원단 음성 신호의 주파수 밴드별 음성 부재 확률 (b) 기존의 음성 강화의 강화 이득 값 (c) 제안된 음성 강화 이득 이득 값

Fig. 3. The comparison of the speech reinforcement gain values. (a) The SAP of the noisy far-end speech signal (b) The gain values of the conventional speech reinforcement algorithm (c) The gain values of the proposed robust speech reinforcement approach based on soft decision.

단되는 구간에서는 음성 신호가 증폭되지 않아야 하지만, 기존의 방법의 경우 강화 이득 값이 1보다 큰 이득 값 2에 가까운 것을 볼 수 있다. 그러나 제안된 방법의 경우에는 오염된 음성 신호로부터 깨끗한 음성 신호의 추정뿐만 아니라, 강화 이득 값에도 soft decision을 적용하였기 때문에, 그림 (c)에서 보는 바와 같이 음성 부재 확률이 1에 가까운 구간에서는 강화 이득 값이 1에 가깝기 때문에 음성의 명료성을 저하시키는 잡음 신호의 증폭이 이루어지지 않음으로써, 기존의 방법 보다 강인한 음성 강화가 이루어짐을 볼 수 있다.

최종 강화 이득 값에 의해 오염된 원단 음성 신호의 DFT 계수는 다음과 같이 나타낼 수 있다.

$$\hat{Y}(t,k) = \tilde{G}_{total}(t,k)\,Y(t,k). \tag{20}$$

그림 4에서는 원단의 음성 신호가 근단의 배경 잡음에 의해 마스킹 되었을 때의 경우로써, (a)는 비교 대상이 되는 음성 파일을 나타내고, (b)에는 제안된 알고리즘이 적용된 음성 파일이다. 그림에서 보는 바와 같이, (a)의 경우 배경 잡음이 부가된 음성 신호는 잡음에 의해 심하게 마스킹 되지만, (b)의 음성 신호는 제안된 음성 강화 이득 수정 알고리즘이 적용되어 잡음 신호 증폭 없이 잡음이 섞이지 않은 깨끗한 음성 신호만을 추정하여 증폭함으로

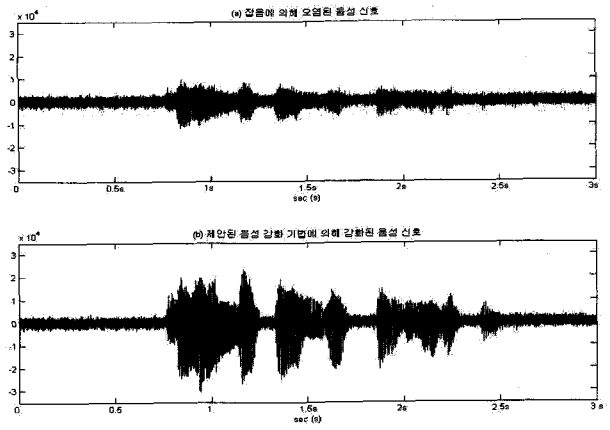

그림 4. 음성 신호의 비교 (a) 잡음에 의해 오염된 원단 음성 신호 (b) 제안된 음성 강화 기법에 의해 강화된 음성 신호

Fig. 4. Comparisons of the speech signal (a) Noisy far-end speech signal (b) The speech signal based on the reinforcement approach.

써, 근단의 다양한 배경 잡음 환경에서 음성의 명료도를 보다 향상시킴을 볼 수 있다.

IV. 실험결과 분석 및 비교

본 논문에서는 제안된 음성 강화 알고리즘의 성능을 평가하기 위해서, 다양한 배경 잡음 환경에서 ITU-T P.800의 주관적인 음질 측정 방법인 Comparison Category Rating (CCR)로 실험을 진행하였다[9]. 먼저, NTT 한국어 음성 데이터베이스에서 20세에서 35세 사이의 남녀 각각 4명씩의 음성 샘플을 추출하였으며, 음성 샘플의 총 길이는 8초이고, 8 kHz로 샘플링 하였다. 원단의 오염된 입력 신호를 위해서, white gaussian noise (WGN)을 SNR 10 dB로 부가 하였다. 근단의 다양한 배경 잡음을 위해서, white, babble, 그리고 vehicle 잡음이 SNR 5, 10, 15, 20 dB로 섞이도록 하였다[6].

총 14명의 한국인 청자에 의해서 주관적 음질 측정 방법인 CCR 테스트를 진행하였으며, 14명의 청자 중에서 8명의 남자이고, 6명은 여성으로 구성되게 하였다. 주관적 음질 측정 방법인 CCR 테스트는 두 음성 신호를 비교하여, 선호도를 평가하는 방법으로써, 비교 대상이 되는 두 신호중 어느 음성 신호가 더 음질이 좋은지, 배경 잡음에 대해서 음성 신호가 얼마나 더 명료하게 잘 들리는지, 얼마나 크게 들리는지를 종합적으로 판단하여 -3점부터 +3점까지 점수를 부여하는 방법이다[9]. 따라서 만약 A와 B

표 1. 다양한 배경 잡음 환경에서 제안된 음성 강화 이득 수정 기법 대비 기존의 SNR 복구 기법 (Frequency Independent SNR Recovery)에 대한 CCR 테스트 결과 (95% 신뢰 구간)

Table 1. The CCR results for the proposed reinforced algorithm with respect to the Frequency Independent SNR Recovery algorithm under various background noise (with 95% confidence interval).

noise	white	babble	vehicle
5 dB	0.708±0.11	0.594±0.13	0.521±0.13
10 dB	0.677±0.11	0.708±0.10	0.531±0.13
15 dB	0.698±0.11	0.750±0.10	0.625±0.12
20 dB	0.656±0.10	0.677±0.10	0.563±0.12
average	0.680±0.11	0.682±0.11	0.560±0.13

표 2. 다양한 배경 잡음 환경에서 제안된 음성 강화 이득 수정 기법 대비 기존의 음성 강화 기법에 대한 CCR 테스트 결과 (95% 신뢰 구간)

Table 2. The CCR results for the proposed reinforced algorithm with respect to the conventional speech reinforcement algorithm under various background noise (with 95% confidence interval).

noise	white	babble	vehicle
5 dB	0.875±0.10	0.844±0.12	0.646±0.14
10 dB	0.771±0.14	0.781±0.11	0.698±0.14
15 dB	0.823±0.12	0.833±0.11	0.688±0.14
20 dB	0.865±0.12	0.844±0.09	0.688±0.14
average	0.834±0.12	0.826±0.11	0.680±0.14

두 음성 신호를 비교해서, A 음성 신호가 좋다면 결과에 나타난 총 점수의 평균값은 0보다 큰 양수 값을 갖게 되며, B의 음성 신호가 더 선호된다면 0보다 작은 음수의 평균값을 가지게 된다.

첫 번째 실험은 다양한 근단의 배경 잡음 환경을 고려하여, 제안된 음성 강화 알고리즘과 기존의 Sauert와 Vary가 제안한 SNR 복구 기법과의 비교를 통해 음성의 명료도를 평가하기 위한 것이다. 표 1에 나타난 결과를 비교해 보면, white 잡음에 대해 95% 신뢰 구간 (=±0.11)에서는 평균 점수 0.680, babble 잡음의 경우 95% 신뢰 구간 (=±0.11)에서는 평균 점수 0.682 및 vehicle 잡음에 대해 95% 신뢰 구간 (=±0.13)에서는 평균 점수 0.560을 나타내었다. 평균 점수가 0보다 크다는 의미는 제안된 음성 강화 기법이 기존의 SNR 복구 기법과 비교하여 근단의 배경 잡음 환경에서 실험에 참가한 청자에 의해서 더 선호된다는 것을 나타내고, 실험의 결과를 통해 제안된 soft decision 통합 음성 강화 이득 수정 알고리즘이 원단 음성 신호의 명료도를 향상시켰다는 것을 알 수 있다.

두 번째 실험은 기존의 오염된 원단 음성 신호로부터 soft decision에 기반하여 깨끗한 음성 신호만을 추정하고, 추정된 깨끗한 음성 신호만을 증폭시키는 알고리즘과 깨끗한 음성 신호의 추정뿐만 아니라, 음성 강화 이득 값 추정을 위하여 제안된 soft decision을 통합한 음성 강화 이득 수정 알고리즘을 비교한 것이다. 표 2의 결과에 따르면, white 잡음의 경우 95% 신뢰 구간 (=±0.12)에서 평균 점수 0.834, babble 잡음에 대해 95% 신뢰 구간 (=±0.11)에서 평균 점수 0.826 및 vehicle 잡음의 경우 95% 신뢰 구간 (=±0.14)에서 평균 점수 0.680을 보여줌으로써 제안된 soft decision을 통합한 음성 강화 알고리즘이 더 우수한 결과를 보여주었다. 표의 결과가 일관된 결과를 나타내지 못하는 것은 잡음의 주파수 성격에 따라 음성 강화의 효과가 다른 영향을 미치기 때문이라 추정된다.

V. 결 론

본 논문에서는 다양한 배경 잡음 환경에서, 오염된 음성 신호를 강화하는 효과적인 음성 강화 이득 수정 알고리즘을 제안하였다. 배경 잡음에 의해 오염된 원단의 음성 신호를 강화하기 위해서, soft decision에 기반한 통계적 신뢰성을 갖는 주파수 채널별 음성 부재 확률 (SAP)을 잡음이 섞이지 않은 깨끗한 음성 신호의 추정뿐만 아니라, 보다 강인한 강화 이득 값의 추정에 주파수 채널별 음성 부재 확률을 적용함으로써, 음성 구간과 비음성 구간 및 음성에서 비음성 구간으로의 전이구간에서 보다 강인한 음성 강화 이득을 추정하였다.

주관적 음질 측정 방법인 CCR 테스트를 통하여 기존의 음성 강화 기법보다 주파수 채널별 음성 부재 확률을 음성 강화 이득에 적용한 결과 white 잡음의 경우 95% 신뢰 구간 (=±0.12)에서 평균 점수 0.834, babble 잡음에 대해 95% 신뢰 구간 (=±0.11)에서 평균 점수 0.826 및 vehicle 잡음의 경우 95% 신뢰 구간 (=±0.14)에서 평균 점수 0.680의 향상된 결과를 보여주었다.

참 고 문 헌

[1] N. S. Kim, J.-H. Chang, "Spectral enhancement

based on global soft decision," IEEE Signal Processing Letters, vol. 7, No. 5, May 2000, pp. 108-110.
[2] J.-H. Chang, N. S. Kim, "Speech enhancement: new approaches to soft decision," IEICE Trans. Inf. and Syst., vol E84-D, pp. 1231-1240, Sep. 2001.
[3] J.-H. Chang, S. Gazor, N. S. Kim and S. K. Mitra, "Multiple statistical models for soft decision in noisy speech enhancement," Pattern Recognition, vol. 40, no. 3, pp. 1123-1134, Mar. 2007.
[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acous., Speech, Signal Process., vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984.
[5] J.-H. Chang, N. S. Kim and S. K. Mitra, "Voice activity detection based on multiple statistical models," IEEE Trans. Signal Processing, vol. 56, no. 6, pp. 1965-1976, June 2006.
[6] J. W. Shin and N. S. Kim, "Perceptual reinforcement of speech signal based on partial specific loudness," IEEE Signal Process. Lett., vol. 14, no. 11, pp. 887-890, Nov. 2007.
[7] B. Sauert and P. Vary, "Near end listening enhancement:Speech intelligibility improvement in noisy environments," in Proc. IEEE Int. Conf. Acoustics., Speech, Signal Processing., vol. 1, pp. I-493-I-496, 2006.
[8] 최재훈, 장준혁, "원단 잡음 환경에서 Soft Decision에 기반한 새로운 음성 강화 기법," 한국음향학회지, 제 27권, 제 7호, pp. 379-385, 2008년 10월.
[9] ITU-T P.800, Methods for Subjective Determination of Transmission Quality, Aug. 1996.

저 자 소 개

최 재 훈(학생회원)
2007년 인하대학교 전자전기 공학부 학사.
2008년 삼성전자 정보통신 총괄 연구원.
2008년~현재 인하대학교 전자공학과 석사과정.
<주관심분야 : 디지털 음성 신호처리>

장 준 혁(정회원)
1998년 경북대학교 전자공학과 학사.
2000년 서울대학교 전기공학부 석사.
2004년 서울대학교 전기컴퓨터 공학부 박사.
2000년~2005년 (주)넷더스 연구소장
2004년~2005년 캘리포니아 주립대학, 산타바바라(UCSB) 박사후연구원
2005년 한국과학기술연구원(KIST) 연구원
2005년~현재 인하대학교 전자공학부 조교수
<주관심분야 : 음성 신호처리, 오디오 신호처리, 통신 신호처리, 휴먼/컴퓨터 인터페이스>